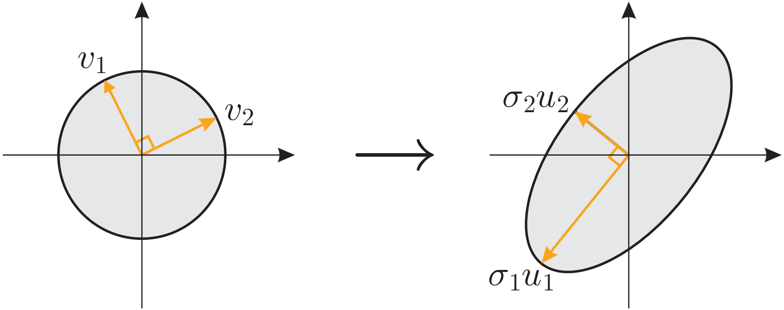
Singular Value Decomposition

Stephen Boyd and Sanjay Lall

EE263
Stanford University

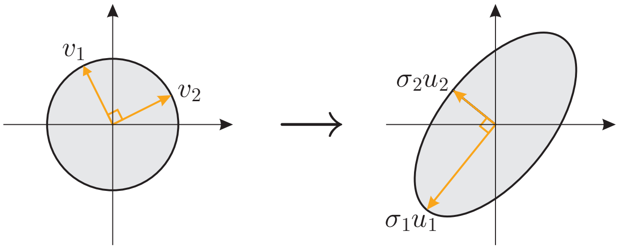
Geometry of linear maps



every matrix $A \in \mathbb{R}^{m \times n}$ maps the unit ball in $\mathbb{R}^n$ to an ellipsoid in $\mathbb{R}^m$

$$S = \left\{ x \in \mathbb{R}^n \mid \|x\| \leq 1 \right\} \qquad AS = \left\{ Ax \mid x \in S \right\}$$

Singular values and singular vectors



- first, assume $A \in \mathbb{R}^{m \times n}$ is skinny and full rank
- the numbers $\sigma_1, \ldots, \sigma_n > 0$ are called the *singular values* of A
- the vectors $u_1, \ldots, u_n$ are called the *left* or *output singular vectors* of A. These are *unit vectors* along the principal semiaxes of AS
- the vectors $v_1, \ldots, v_n$ are called the *right* or *input singular vectors* of A.
  These map to the principal semiaxes, so that

$$Av_i = \sigma_i u_i$$

Thin singular value decomposition

$$Av_i = \sigma_i u_i \qquad \text{for } 1 \le i \le n$$

For $A \in \mathbb{R}^{m \times n}$ with $\mathbf{rank}(A) = n$, let

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \qquad V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

the above equation is $AV = U\Sigma$ and since V is orthogonal

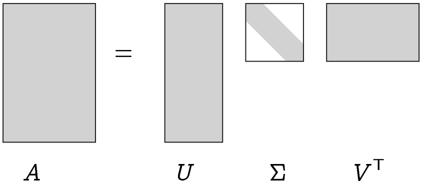
$$A = U\Sigma V^{\mathsf{T}}$$

called the *thin SVD* of A

Thin SVD

For $A \in \mathbb{R}^{m \times n}$ with $\mathbf{rank}(A) = r$, the *thin SVD* is

$$A = U\Sigma V^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$



here

- $U \in \mathbb{R}^{m \times r}$ has orthonormal columns,
- $\Sigma = \mathbf{diag}(\sigma_1, \dots, \sigma_r)$, where $\sigma_1 \geq \cdots \geq \sigma_r > 0$
- $V \in \mathbb{R}^{n \times r}$ has orthonormal columns

SVD and eigenvectors

$$A^\top A = (U\Sigma V^\top)^\top (U\Sigma V^\top) = V\Sigma^2 V^\top$$

hence:

- v_i are eigenvectors of $A^\top A$ (corresponding to nonzero eigenvalues)
- $\sigma_i = \sqrt{\lambda_i(A^\top A)}$ (and $\lambda_i(A^\top A) = 0$ for $i > r$)
- $||A|| = \sigma_1$

SVD and eigenvectors

similarly,

$$AA^\top = (U\Sigma V^\top)(U\Sigma V^\top)^\top = U\Sigma^2 U^\top$$

hence:

- u_i are eigenvectors of $AA^\top$ (corresponding to nonzero eigenvalues)
- $\sigma_i = \sqrt{\lambda_i(AA^\top)}$ (and $\lambda_i(AA^\top) = 0$ for $i > r$)

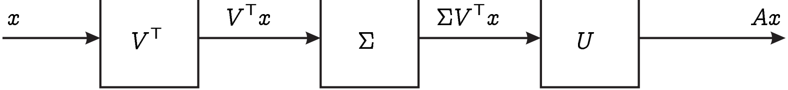
SVD and range

$$A = U\Sigma V^\top$$

- $u_1, \ldots u_r$ are orthonormal basis for **range**(A)
- $v_1, \ldots v_r$ are orthonormal basis for **null**$(A)^\perp$

Interpretations

$$A = U\Sigma V^\mathsf{T} = \sum_{i=1}^{r} \sigma_i u_i v_i^\mathsf{T}$$



linear mapping $y = Ax$ can be decomposed as

- compute coefficients of x along input directions $v_1, \ldots, v_r$
- scale coefficients by σ_i
- reconstitute along output directions $u_1, \ldots, u_r$

difference with eigenvalue decomposition for symmetric A: input and output directions are *different*

Gain

- v_1 is most sensitive (highest gain) input direction
- u_1 is highest gain output direction
- $Av_1 = \sigma_1 u_1$

Gain

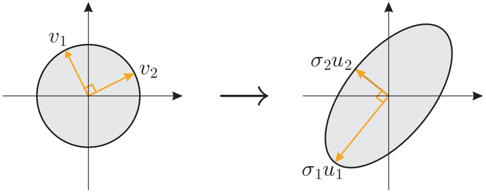
SVD gives clearer picture of gain as function of input/output directions

example: consider $A \in \mathbb{R}^{4\times 4}$ with $\Sigma = \mathbf{diag}(10,\ 7,\ 0.1,\ 0.05)$

- input components along directions v_1 and v_2 are amplified (by about 10) and come out mostly along plane spanned by u_1, u_2
- input components along directions v_3 and v_4 are attenuated (by about 10)
- $||Ax||/||x||$ can range between 10 and 0.05
- A is nonsingular
- for some applications you might say A is *effectively* rank 2

Example: SVD and control

we want to choose x so that $Ax = y_{\text{des}}$.



- right singular vector v_i is mapped to left singular vector u_i, amplified by σ_i
- σ_i measures the *actuator authority* in the direction $u_i \in \mathbb{R}^m$
- $r < m \quad \Longrightarrow$ no control authority in directions $u_{r+1}, \ldots, u_m$
- if A is fat and full rank, then the ellipsoid is

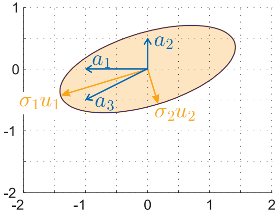
$$E = \left\{ y \in \mathbb{R}^m \mid y^\mathsf{T} \left(AA^\mathsf{T}\right)^{-1} y \leq 1 \right\}$$

because

$$AA^\mathsf{T} = U\Sigma V^\mathsf{T} V \Sigma U^\mathsf{T} = U\Sigma^2 U^\mathsf{T}$$

Example: Forces applied to a rigid body

apply forces via thrusters x_i in specific directions



$$A = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 0.5 & -0.5 \end{bmatrix}$$

- total force on body $y = Ax$,
- x_i is power (in W) supplied to thruster i
- $||a_i||$ is *efficiency* of thruster
- most efficient direction we can apply thrust is given by long axis
- $\sigma_1 = 1.4668$, $\sigma_2 = 0.5904$

General pseudo-inverse

if $A \neq 0$ has SVD $A = U\Sigma V^T$, the *pseudo-inverse* or *Moore-Penrose inverse* of A is

$$A^\dagger = V\Sigma^{-1}U^T$$

- if A is skinny and full rank,
$$A^\dagger = (A^TA)^{-1}A^T$$
gives the least-squares approximate solution $x_{\rm ls} = A^\dagger y$
- if A is fat and full rank,
$$A^\dagger = A^T(AA^T)^{-1}$$
gives the least-norm solution $x_{\rm ln} = A^\dagger y$

General pseudo-inverse

$$X_{\mathrm{ls}} = \{ \ z \mid \|Az - y\| = \min_w \|Aw - y\| \ \}$$

is set of least-squares approximate solutions

$x_{\mathtt{pinv}} = A^\dagger y \in X_{\mathrm{ls}}$ has minimum norm on X_{ls}, *i.e.*, $x_{\mathtt{pinv}}$ is the minimum-norm, least-squares approximate solution

Pseudo-inverse via regularization

for $\mu > 0$, let x_μ be (unique) minimizer of

$$\|Ax - y\|^2 + \mu\|x\|^2$$

$i.e.,$

$$x_\mu = \left(A^T A + \mu I\right)^{-1} A^T y$$

here, $A^T A + \mu I > 0$ and so is invertible

then we have $\lim_{\mu \to 0} x_\mu = A^\dagger y$

in fact, we have $\lim_{\mu \to 0} \left(A^T A + \mu I\right)^{-1} A^T = A^\dagger$

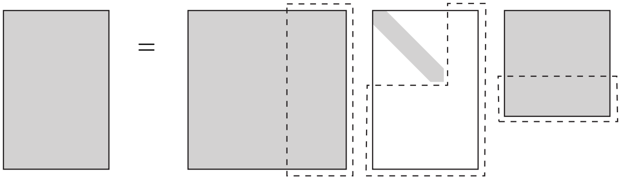
(check this!)

Full SVD

SVD of $A \in \mathbb{R}^{m \times n}$ with **rank**$(A) = r$

$$A = U_1 \Sigma_1 V_1^\mathsf{T} = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^\mathsf{T} \\ \vdots \\ v_r^\mathsf{T} \end{bmatrix}$$

Add extra columns to U and V, and add zero rows/cols to Σ_1



A $\quad U \quad \Sigma \quad V^\mathsf{T}$

Full SVD

- find $U_2 \in \mathbb{R}^{m\times(m-r)}$ such that $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{m\times m}$ is orthogonal
- find $V_2 \in \mathbb{R}^{n\times(n-r)}$ such that $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \in \mathbb{R}^{n\times n}$ is orthogonal
- add zero rows/cols to Σ_1 to form $\Sigma \in \mathbb{R}^{m\times n}$

$$\Sigma = \left[\begin{array}{c|c} \Sigma_1 & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right]$$

then the full SVD is

$$A = U_1\Sigma_1 V_1^\mathsf{T} = \left[\begin{array}{c|c} U_1 & U_2 \end{array}\right] \left[\begin{array}{c|c} \Sigma_1 & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right] \left[\begin{array}{c} V_1^\mathsf{T} \\ \hline V_2^\mathsf{T} \end{array}\right]$$

which is $A = U\Sigma V^\mathsf{T}$

example: SVD

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 4 & 2 \end{bmatrix}$$

SVD is

$$A = \begin{bmatrix} -0.319 & 0.915 & -0.248 \\ -0.542 & -0.391 & -0.744 \\ -0.778 & -0.103 & 0.620 \end{bmatrix} \begin{bmatrix} 5.747 & 0 \\ 0 & 1.403 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.880 & -0.476 \\ -0.476 & 0.880 \end{bmatrix}$$

Image of unit ball under linear transformation

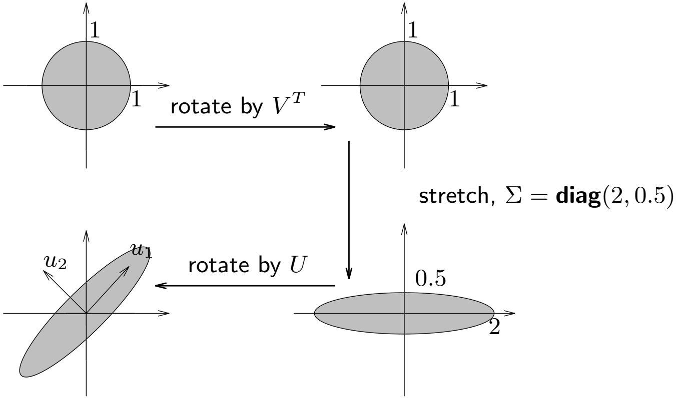
full SVD:

$$A = U\Sigma V^\top$$

gives intepretation of $y = Ax$:

- rotate (by $V^\top$)
- stretch along axes by σ_i ($\sigma_i = 0$ for $i > r$)
- zero-pad (if $m > n$) or truncate (if $m < n$) to get m-vector
- rotate (by U)

Image of unit ball under A



$\{Ax \mid \|x\| \leq 1\}$ is *ellipsoid* with principal axes $\sigma_i u_i$.

Sensitivity of linear equations to data error

consider $y = Ax$, $A \in \mathbb{R}^{n \times n}$ invertible; of course $x = A^{-1}y$

suppose we have an error or noise in y, $i.e.$, y becomes $y + \delta y$

then x becomes $x + \delta x$ with $\delta x = A^{-1}\delta y$

hence we have $||\delta x|| = ||A^{-1}\delta y|| \leq ||A^{-1}|| ||\delta y||$

if $||A^{-1}||$ is large,

- small errors in y can lead to large errors in x
- can't solve for x given y (with small errors)
- hence, A can be considered singular in practice

Relative error analysis

a more refined analysis uses *relative* instead of *absolute* errors in x and y

since $y = Ax$, we also have $||y|| \leq ||A||||x||$, hence

$$\frac{||\delta x||}{||x||} \leq ||A||||A^{-1}||\frac{||\delta y||}{||y||}$$

So we define the *condition number* of A:

$$\kappa(A) = ||A||||A^{-1}|| = \sigma_{\mathsf{max}}(A)/\sigma_{\mathsf{min}}(A)$$

Relative error analysis

we have:

$$\text{relative error in solution } x \leq \text{condition number} \cdot \text{relative error in data } y$$

or, in terms of # bits of guaranteed accuracy:

$$\# \text{ bits accuracy in solution} \approx \# \text{ bits accuracy in data} - \log_2 \kappa$$

we say

- A is well conditioned if κ is small
- A is poorly conditioned if κ is large

(definition of 'small' and 'large' depend on application)

same analysis holds for least-squares approximate solutions with A nonsquare, $\kappa = \sigma_{\max}(A)/\sigma_{\min}(A)$

Low rank approximations

suppose $A \in \mathbb{R}^{m \times n}$, **rank**$(A) = r$, with SVD $A = U\Sigma V^\mathsf{T} = \sum_{i=1}^{r} \sigma_i u_i v_i^\mathsf{T}$

we seek matrix $\hat{A}$, **rank**$(\hat{A}) \le p < r$, s.t. $\hat{A} \approx A$ in the sense that $\|A - \hat{A}\|$ is minimized

solution: optimal rank p approximator is

$$\hat{A} = \sum_{i=1}^{p} \sigma_i u_i v_i^\mathsf{T}$$

- hence $\|A - \hat{A}\| = \left\| \sum_{i=p+1}^{r} \sigma_i u_i v_i^\mathsf{T} \right\| = \sigma_{p+1}$
- interpretation: SVD dyads $u_i v_i^\mathsf{T}$ are ranked in order of 'importance'; take p to get rank p approximant

Proof: Low rank approximations

suppose **rank**$(B) \leq p$

then **dim null**$(B) \geq n - p$

also, **dim span**$\{v_1, \ldots, v_{p+1}\} = p + 1$

hence, the two subspaces intersect, $i.e.$, there is a unit vector $z \in \mathbb{R}^n$ s.t.

$$Bz = 0, \qquad z \in \textbf{span}\{v_1, \ldots, v_{p+1}\}$$

$$(A - B)z = Az = \sum_{i=1}^{p+1} \sigma_i u_i v_i^\mathsf{T} z$$

$$\|(A - B)z\|^2 = \sum_{i=1}^{p+1} \sigma_i^2 (v_i^\mathsf{T} z)^2 \geq \sigma_{p+1}^2 \|z\|^2$$

hence $\|A - B\| \geq \sigma_{p+1} = \|A - \hat{A}\|$

Distance to singularity

another interpretation of σ_i:

$$\sigma_i = \textbf{min}\{ \|A - B\| \mid \textbf{rank}(B) \leq i - 1 \}$$

$i.e.$, the distance (measured by matrix norm) to the nearest rank $i - 1$ matrix

for example, if $A \in \mathbb{R}^{n \times n}$, $\sigma_n = \sigma_{\textbf{min}}$ is distance to nearest singular matrix

hence, small $\sigma_{\textbf{min}}$ means A is near to a singular matrix

Application: model simplification

suppose $y = Ax + v$, where

- $A \in \mathbb{R}^{100\times 30}$ has singular values

$$10,\ 7,\ 2,\ 0.5,\ 0.01,\ \ldots, 0.0001$$

- $||x||$ is on the order of 1
- unknown error or noise v has norm on the order of 0.1

then the terms $\sigma_i u_i v_i^\top x$, for $i = 5, \ldots, 30$, are substantially smaller than the noise term v

simplified model:

$$y = \sum_{i=1}^{4} \sigma_i u_i v_i^\top x + v$$

Example: Low rank approximation

$$A = \begin{bmatrix} 11.08 & 6.82 & 1.76 & -6.82 \\ 2.50 & -1.01 & -2.60 & 1.19 \\ -4.88 & -5.07 & -3.21 & 5.20 \\ -0.49 & 1.52 & 2.07 & -1.66 \\ -14.04 & -12.40 & -6.66 & 12.65 \\ 0.27 & -8.51 & -10.19 & 9.15 \\ 9.53 & -9.84 & -17.00 & 11.00 \\ -12.01 & 3.64 & 11.10 & -4.48 \end{bmatrix}$$

$$\approx \begin{bmatrix} -0.25 & 0.45 & 0.62 & 0.33 & 0.46 & 0.05 & -0.19 & 0.01 \\ 0.07 & 0.11 & 0.28 & -0.78 & -0.10 & 0.33 & -0.42 & 0.05 \\ 0.21 & -0.19 & 0.49 & 0.11 & -0.47 & -0.61 & -0.24 & -0.01 \\ -0.08 & -0.02 & 0.20 & 0.06 & -0.27 & 0.30 & 0.20 & -0.86 \\ 0.50 & -0.55 & 0.14 & -0.02 & 0.61 & 0.02 & -0.08 & -0.20 \\ 0.44 & 0.03 & -0.05 & 0.50 & -0.30 & 0.55 & -0.36 & 0.18 \\ 0.59 & 0.43 & 0.21 & -0.14 & -0.03 & -0.00 & 0.62 & 0.13 \\ -0.30 & -0.51 & 0.43 & 0.02 & -0.14 & 0.34 & 0.41 & 0.40 \end{bmatrix} \begin{bmatrix} 36.83 & 0 & 0 & 0 \\ 0 & 26.24 & 0 & 0 \\ 0 & 0 & 0.02 & 0 \\ 0 & 0 & 0 & 0.01 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.04 & -0.54 & -0.61 & 0.58 \\ 0.92 & 0.17 & -0.33 & -0.14 \\ -0.14 & -0.49 & -0.31 & -0.80 \\ -0.36 & 0.66 & -0.65 & -0.09 \end{bmatrix}$$

$$A_{\text{approx}} \approx \begin{bmatrix} -0.25 & 0.45 \\ 0.07 & 0.11 \\ 0.21 & -0.19 \\ -0.08 & -0.02 \\ 0.50 & -0.55 \\ 0.44 & 0.03 \\ 0.59 & 0.43 \\ -0.30 & -0.51 \end{bmatrix} \begin{bmatrix} 36.83 & 0 \\ 0 & 26.24 \end{bmatrix} \begin{bmatrix} -0.04 & -0.54 & -0.61 & 0.58 \\ 0.92 & 0.17 & -0.33 & -0.14 \end{bmatrix}$$

Example: Low rank approximation

$$A = \begin{bmatrix} 11.08 & 6.82 & 1.76 & -6.82 \\ 2.50 & -1.01 & -2.60 & 1.19 \\ -4.88 & -5.07 & -3.21 & 5.20 \\ -0.49 & 1.52 & 2.07 & -1.66 \\ -14.04 & -12.40 & -6.66 & 12.65 \\ 0.27 & -8.51 & -10.19 & 9.15 \\ 9.53 & -9.84 & -17.00 & 11.00 \\ -12.01 & 3.64 & 11.10 & -4.48 \end{bmatrix}$$

$$A_{\text{approx}} = \begin{bmatrix} 11.08 & 6.83 & 1.77 & -6.81 \\ 2.50 & -1.00 & -2.60 & 1.19 \\ -4.88 & -5.07 & -3.21 & 5.21 \\ -0.49 & 1.52 & 2.07 & -1.66 \\ -14.04 & -12.40 & -6.66 & 12.65 \\ 0.27 & -8.51 & -10.19 & 9.15 \\ 9.53 & -9.84 & -17.00 & 11.00 \\ -12.01 & 3.64 & 11.10 & -4.47 \end{bmatrix}$$

here $||A - A_{\text{approx}}|| \leq \sigma_3 \approx 0.02$